The practicality of any nonparametric statistical procedure should be confirmed thoroughly with regard to the data distribution under study

Maciej Górkiewicz[1], **Agnieszka Gniadek**[2]

[1] Department of Epidemiology and Population Research, Jagiellonian University Medical College, Poland

[2] Department of Medical and Environmental Nursing, Faculty of Health Sciences, Jagiellonian University Medical College, Poland

Abstract. The present study was motivated by pilot research aimed to examine the aptness of the anti-fungal everyday activity in the hospital settings. The number of uncovered fungi colonies per cubic meter of air was chosen here as the crucial indicator of the quality. The empirical probability distribution functions of this indicator at various hospital's wards showed great variety of their shapes. Nevertheless, from the practical point of view, preferably the comparisons between these functions should be expressed in terms of the mean values. Therefore, here the practical problem arose: how to avoid inappropriate choice of a nonparametric statistical test for equivalence of the mean values given random data sets. In the present study, the review of the nonparametric methods was limited to the most popular ones only: the Box-Cox transformation, the Mann-Whitney rank sum test, and the log-rank approach. The more advanced formal considerations were omitted. The limitations of the Box-Cox transformation and the Mann-Whitney rank sum test were explained with clear examples based on the artificial samples. Practical criterions, helpful to avoid common pitfalls and misunderstandings were recommended. The advantages and the weaknesses of the log-rank approach were demonstrated basing on the real-life data sets.

Introduction

Any statistical procedure was founded on a number of assumptions regarding not only formal features of the data under investigation, but also the attributes of the anticipated area of application of the results of the statistical analysis. In the real world, the departures from these ideal assumptions are unavoidable [1–3]. Thus, in applied research the practicality of the results has two faces. First, the all way of analyses should be correct from a pure mathematical point of view, and then, from the practical perspective, the real meaning of the results should correspond with the anticipated area of their use.

ISBN 978–83–7431–336–0 ISSN 0860-150X

The parametric procedures were constructed under common assumption that the data samples were drawn from a normal distribution. For that reason, if the empirical distribution of the data distinctly differs from ideal normal distribution, then some nonparametric procedure is usually applied. Unfortunately, many non-statisticians are wrongly convinced, that in such a situation, the use of the most suitable nonparametric "ersatz" doesn't change the essence of the conclusions from the calculations [4–5]. For instance, numerous non-statisticians wrongly believe that parametric Student t-test, applied to the data after Box-Cox transformation, gives conclusion regarding data before transformation. Many other non-statisticians accept as true the imprecise supposition that the Mann-Whitney rank sum test examines the relation between medians. On the other side, it is known that the parametric procedures are fairly robust to the moderate departures from normality [6], at least the obtained results need some moderate corrections with regard to estimated skew and kurtosis coefficients [7]. In consequence, other numerous non-statisticians, dealing with the reasonable number of random data, do not take into account the use of any nonparametric procedure. They wrongly believe that for any kind of distributions, the parametric procedures at any circumstances generate reliable conclusions, at least with regard to the relation between the mean values.

All above mistaken beliefs can lead to misinterpretation of the data under examination, subsequently to wrong practical decisions, and as a final result, to depreciation of the statistical methodology in the public opinion. In the literature, the discussed potential causes of this undesirable phenomenon include:

(i) pressure 'publish or perish' on candidate researchers [8];
(ii) wishful thinking instead the critical one [9–10];
(iii) not-user-friendly style and confusing terminology applied in the statistical textbooks, like the misleading idiom 'distribution free' frequently used with regard to some nonparametric procedure [11–12];
(iv) numerous silent assumptions in the statistical instructions, that are obvious for the statisticians, but rather hard to reveal for the others [13–14].

Several fundamental changes in statistical training and practice are recommended in literature, with a general purpose of changing for the better of this situation. It was suggested to make a greater emphasis on the philosophy underlying the statistical methodology [15], with special focus on a common-sense approach and on the possible pitfalls and misinterpretations [16–18]. The wide-ranging use of the exploratory data analysis is postulated, first before starting the usual confirmatory data analysis, with the aim to reveal unexpected or misleading patterns in the data and to foster hypothesis development and refinement, and then after this, with the aim to help one interpret the obtained results [19]. Modern tools for the visual exploration of large databases create an opportunity to discover the outliers and clusters within the data [20] and to confirm the fit to supposed distribution [21], avoiding the false impressions caused by inspection limited to the traditional histograms only [22].

The present study corresponds in general with all the above cited ideas concerning the desirable improvements in the current research practice, but the sphere of interest was strictly limited here to a practical problem: how to assess the aptness of anti-fungal everyday prevention. Consequently, the rest of this paper was organized as follows. First, the problem how to assess the quality of the everyday anti-fungal clinical practice is discussed. This section can be omitted by person non interested in the clinical problems. The review of the nonparametric methods started with the most popular transformations aimed to diminish the influence of non-normality. The limitation of the Box-Cox method was shown on the exemplary data, and some other methods were briefly characterized on the base of literature. Then, the Wilcoxon-Mann-Whitney rank-sum test was examined in aspect to the question if the results of this test can be interpreted in terms of medians and usual arithmetic means. In the next section, the descriptive statistics for the motivating practical question was made in two tables including the main characteristic of the $2 \cdot (45 + 20) = 130$ samples obtained in patients' rooms at two hospital wards. Finally, the classical log-rank methodology was applied in the study. More advanced permutation tests and stochastic modelling procedures were only discussed as the possible subject for further investigations.

How to assess the aptness of clinical everyday anti-fungal practice?

In clinical settings, the primary mode of acquiring a mould infection is inhalation of room air polluted with fungal spore-loaded dust [23]. Therefore, the concentration of the fungi in the air, that is the number of uncovered fungi colonies per cubic meter of air, at various hospital wards was commonly acknowledged as the crucial indicator of the quality of the overall anti-fungal activity [24]. The appropriate isolation of patients from harmful aerosols might be achieved only with combined use of several means [25]. Moreover, the final effectiveness of usual everyday antifungal activity must be under permanent inspection with the aim to put into action routines, like wearing of filtering masks [26]. This is avery challenging problem, from practical as well as theoretical point of view, because the anti-fungal activity represents only a component of the whole very complex system named – good clinical practice. Therefore, there arises a great difficulty in defining elements of the intervention. It may be impossible to single out which particular parts are effective, since one component may not work without another [27–28].

Therefore, in our best knowledge, the problem, how to utilize the measurements of the fungi concentration at the given hospital with the aim to improve the current system of the anti-fungal activity, up to now haven't any acknowledged practicable solution. The general idea of the our solution to this problem is as follows. The anti-fungal practice, carried out at each particular hospital, should be considered primarily at the whole, as a complex intervention, observing the *primum non nocere* principle [29–31]. Initially, it can be advantageous to consider the relatively simple question: are the frequencies of departures of the concentration of the fungi in the air over appropriate levels in the patients' rooms at the hospital under examination are at least as low as the frequencies in the analogous wards at other good hospitals? After that, before starting with any serious modifications, one should get the reliable answers to the two main initial questions:

(i) are the mean values of the concentration of the fungi in the air in various wards at the hospital under examination are at least as low as that concentration measured with use of the comparable methodology in the analogous wards at the other good hospitals?

(ii) are the mean values of the concentration of the fungi in the wards at the hospital under examination decreased gradually from the maximal value in the entry room to appropriate values inside, at the patients' rooms?

In our investigations of the fungi concentration in the air at various hospitals wards, it was proved that there wasn't the significant correlation between measurements made in adjacent moments in the sequences of morning-evening measurements. Therefore, the series of the measurements made at the same place can be considered as random samples of the uncorrelated data [32]. Nevertheless, there occurred two interacted difficulties: the underlying distributions are far from normal, so the use of the nonparametric methodology should be considered; nonetheless, from the practical point of view, the conclusions must refer to mean values (that is to the expected values) directly, without any subsidiary substitution with some other attribute, like median, geometric mean, harmonic mean, trimmed mean and so on. In our investigations data sets corresponded to the Weibull distribution. Therefore, the search for the most acceptable method can be limited to the review of the parametric and nonparametric test for comparing means of the Weibull populations [33]. Nevertheless it seems to be more appropriate not to ignore the most popular universal procedures, like Box-Cox transformation and the Wilcoxon-Mann-Whitney rank-sum test.

Transformations aimed to diminish the influence of non-normality

The real-life samples examined in applied research fairly often showed attributes atypical for a normal distribution, like coefficients of skewness and kurtosis far from zero, outliers, heavy tails. Since the consequences of non-normality for test statistics are difficult to investigate, many studies suggested the use of transformation procedures developed for specific forms of non-normality. The Box-Cox method was developed for distributions intermediate between the normal and the log-normal distributions, with the aim to restore normality of the data. The Box-Cox transformation was defined with formulas (1), (2), (3).

$$Y = (X^C - 1)/C; \quad \text{for } C \neq 0; \tag{1}$$

either

$$Y = \ln(X); \qquad \text{for } C = 0; \tag{2}$$

$$\text{optimal}(C) = C|\text{optimal}(J) \tag{3}$$

were:

- X – transformed variable before Box-Cox transformation;
- Y – transformed variable after Box-Cox transformation;
- C – power parameter;
- J – criterion of optimality, assumed correspondingly to anticipated parametric procedure of the further analyses.

[Tab. 1] includes two examples. In the first example, the populations numbered 1, 2, and 3 in [Tab. 1], had the same mean values of variable Y, equal to $m_y = 0$, but different variances V_y. In result, the mean values of variable X in these populations occurred manifestly different, accordingly to known formula (4). In the second example, the exemplary populations numbered 4, 5, and 6 in the [Tab. 1], had the same mean values of variable X, equal to $m_x = 1.65$, but different variances V_y. In result, the mean values of variable Y in these populations occurred manifestly different, accordingly to known formula (4).

$$\ln(m_x) = m_y + V_y/2 \tag{4}$$

where:

m_x – mean value of variable X with log-normal distribution;
m_y, V_y – mean value and variance of variable $Y = \ln(X)$ with normal distribution.

Tab. 1. Relationship between mean values of random variable X with log-normal distribution and variable Y = ln(X) in some exemplary populations

population	1	2	3	4	5	6
$m_y = \ln(Me_x)$	0	0	0	– 0.5	0	0.25
V_y	0.5	1	2	2	1	0.5
m_x	1.28	1.65	2.72	1.65	1.65	1.65

m_x, Me_x – mean value and median of variable X with log-normal distribution;
m_y, V_y – mean value and variance of variable $Y = \ln(X)$ with normal distribution.

It is easy to notice, directly from definition of the log-normal distribution [34], that if the variable X in formula (2) has the ideal log-normal distribution, then the variable Y after Box-Cox transformation obtains the ideal normal distribution. Therefore, without any doubts any parametric procedure may be applied to examine mean values of the variable Y among some populations under examination. Moreover, usually there aren't any serious doubts that the ANOVA parametric procedure can be applied with respect to mean values of the variable Y in random samples drawn from exemplary populations numbered 1, 2, and 3 in [Tab. 1], because ANOVA is fairly robust to moderate heteroscedasticity [6], so use of any known counterpart to ANOVA [35–37], will be to no purpose here. The question is, if the proved relationship between mean values of the variable $Y = \ln(X)$ may be adopted to the actually interested relationship between mean values of the variable X, $X > 0$.

The criterions for the appropriate use of the Box-Cox method can be easily deduced from the examples considered in [Tab. 1]. It was demonstrated that the normality of the variable after Box-Cox transformation doesn't give sufficient reason to make an inference about mean values of this variable before Box-Cox transformation, basing on the relationship among mean values of this variable after Box-Cox transformation, as proved with some parametric test. Thus, except for the normality, the homogeneity of the variances of the variable after Box-Cox transformation must be proved too. If the request of the homogeneity isn't fulfilled than the conclusions relate to some other averages, e.g. to the geometric mean for transformation $Y = \ln(X)$, but not to usual arithmetic mean values. Guo and Luh [38–39], discussed several other transformation, more suitable for non-normal distributions that are affected by heavy tails or outliers. There also the results of a parametric procedure applied correctly to the transformed samples, can be related to some special averages of the variable before transformation, but generally not to the usual arithmetic mean values.

The general conclusion from the above review can be summarised as follows. It can be very advantageous to apply a transformation approach dealing with unusual distributions at pilot studies and exploratory data analyses, because it is a quick and easily computable method. However, it should be implemented with great caution.

Wilcoxon-Mann-Whitney rank-sum test

Wilcoxon-Mann-Whitney rank-sum test, or shortly, Mann-Whitney test, pertains to some statistical comparison of two separate populations given two independent random samples, but it is usually thought as the most reliable nonparametric alternative for 2-sample Student t-test in situations where the data appear to arise from non-normal distributions. For this reason, it is easily available at almost all popular statistical packages. The concise introduction to this test with very intuitive graphics one can find in [40].

Let $X = x$ denotes a single random number drawn from the first population, and $Y = y$ denotes a single random number drawn independently from the other separate population. The Mann-Whitney test has been performed as the non-parametric alternative to the parametric Student t-test, but in essence it examines the null hypothesis (5).

$$H_0: \qquad \Pr(x < y) = \Pr(x > 0) \tag{5}$$

under restriction:

$$\Pr(x = y) = 0 \tag{6}$$

where random variables X and Y are both measured at least on an ordinal scale.

Therefore, many non-statisticians are wrongly convinced that for any distributions of variables X and Y the null hypothesis (5) is perfectly equiponderant with the null hypothesis (7) on equivalence of the medians of these variables, and consequently, that the null hypothesis (5) is perfectly equiponderant with the null hypothesis (8) on equivalence of the mean values of these variables, at least for the symmetrical (non-skewed) distributions.

$$H_0': \qquad Me_x = Me_y \tag{7}$$

$$H_0'': \qquad m_x = m_y \tag{8}$$

with silent (wrong) justification: because hypothesis (5) holds $\Pr(x < y) = \Pr(x > 0)$;
where m_x, Me_x, m_y, Me_y – mean values and medians of variables X, Y respectively.

Both of the above convictions are generally wrong, and in practice often lead either to disadvantageous decisions or at best to absurd conclusions in a particular matter under study. The last statement can be easily supported by a simple example shown in [Tab. 2]. In this table the three exemplary cases were constructed in such a way that the restriction (6) was satisfied in each case. Then, the symmetrical distribution of the variable X, and the shape of the symmetrical distribution of the variable Y, both remain the same at all three exemplary cases, but the distribution of Y is shifted to right in case 2, and is shifted to left in case 3. Therefore, the medians of these distributions, initially the same in case 1, are manifestly different from the others, as well in case 2 as in case 3. In other words, the hypotheses (7) and (8) are fulfilled in case 1 only, but they are manifestly violated in the both two remaining cases. Nevertheless, it is easy to notice that in all three cases the hypothesis (5) is evidently satisfied, because probabilities $\Pr(x < y) = \Pr(x > 0) = \frac{1}{2}$ didn't change from case to case.

Tab. 2. Mann-Whitney test for some exemplary pair of the symmetrical distributions

case	shift	$f(X) > 0$	$m_x = Me_x$	$f(Y) > 0$	$m_y = Me_y$	p_{MW}
1	0	$-1 < X < +1$	0	$Y < -2 \cdot 10^6$ or $Y > +2 \cdot 10^6$	0	0.5
2	$+10^6$	$-1 < X < +1$	0	$Y < -1 \cdot 10^6$ or $Y > +3 \cdot 10^6$	$+10^6$	0.5
3	-10^6	$-1 < X < +1$	0	$Y < -3 \cdot 10^6$ or $Y > +1 \cdot 10^6$	-10^6	0.5

$f(X)$, $f(Y)$ – density of symmetrical distribution of the variable X and Y respectively;
m_x, Me_x, m_y, Me_y – mean values and medians of variables X, Y respectively;
p_{MW} – ideal (expected) value of the significance of the Mann-Whitney rank-sum test.

Let us consider a somewhat more down-to-earth situation of drawing samples from the unknown distributions assumed in the above [Tab. 2]. For instance, let X and Y be an anticipated incomes, expressed in \$, from two kind of businesses. It is easy to notice that in each case in [Tab. 2] the probability $\Pr(y < \min(x)) = \Pr(y > \max(x)) = \frac{1}{2}$. Therefore, if the N random values of the variable Y is drawn independently one from other, then the probability that exactly $N/2$ values of Y will occur beyond $\min(x) = -1$, equal to probability that exactly $N/2$ values of Y will occur over $\max(x) = +1$, will depend only from N, and for instance, for the moderate dimension of a sample, $N = 64$, in average only on one occasion in the 20 experiments the sample will not divided into two exactly the same parts, first one of $N/2 = 32$ x's beyond $\min(x) = -1$, and second one of $N/2 = 32$ x's over $\max(x) = +1$. If the random sample of x's has there a dimension also equal to $N = 32$, then the sums of the ranks of the y's and x's occur the same, equal to $32 \cdot (1 + 128) = 32 \cdot (33 + 96) = 4.128$ as well for variable Y as for variable X. A 'naïve' researcher, believing without any doubts that Mann-Whitney test can be directly related to medians, at almost each time will find a reason for an interpretation that an expected balance equal to zero dollars doesn't differ significantly from an expected income equal to million dollars (case 2 in [Tab. 2]), or that it doesn't differ significantly from an expected loss equal to million dollars (case 3 in [Tab. 2]).

The exemplary distributions of Y, in each case considered in [Tab. 2], are divided into two parts, with a gap of density $f(Y) = 0$ between them. It should be noted that a quite similar manifestation of the relations between hypotheses (5), (7), and (8), can be modelled without this gap, also with distribution with a single mode, under restrictions that ratio $f(Y)/f(X)$ has a pattern either low-high-low or high-low-high, so it isn't an example for the Simpson's paradox [41].

The general conclusion from the above considerations can be summarised as follows. The Mann-Whitney test can be applied without any hesitation to practical problems that can be expressed in terms of the hypothesis (5), without any serious focus on the medians or on the usual arithmetic means. If the problem under study must be related at least to medians, like hypothesis (7), then the ratio $f(Y)/f(X)$ should be proved with respect to its monotonicity. If the problem under study must be related to the usual arithmetic means, then additionally both distributions, the $f(Y)$ and the $f(X)$ should be sincerely symmetrical. As to the last case, it is known, that for symmetrical non-normal distributions, the differences in the power between Student t test and Mann-Witney rank-sum test are so small that the choice is immaterial for practical purposes [42–43].

Motivating example

[Tab. 3] showed the descriptive statistics of the concentration of fungi in the air in patients' rooms at two hospital wards under study, as measured in the morning and in the evening during five consecutive winter days. It is easy to notice the evident departure from normality, with a mean values rather far from the medians, and relative great coefficients as well as for skew as for kurtosis. Therefore, the parametric tests, like ANOVA and Student t-test, seem to be inappropriate here. From practical reasons, the results of comparisons should be related to usual arithmetic means. Thus, the most popular counterparts, like transformations and Mann-Whitney rank-sum test, also seems to be quite inappropriate to apply in the matter.

In such a case, the simulation approach seems to be most suitable [44–45], in particular with respect to easy available on-line calculators [46–47]. On the other hand, in this study the log-rank plots supported supposition that the distribution ofconcentration of the fungi are consistent with Weibull distribution, see [Tab. 4]. Consequently, with the aim to make comparisons between average concentrations of fungi at the different times and sites, the log-rank approach was applied.

Tab. 3. Descriptive statistics of the concentration of fungi in air in the patient's bedrooms during five consecutive days

ID	ward	time	N	mean	SD	median	min.	max	skew	kurtosis
1	HP	morning	45	35.3	47.9	15	0	195	1.78	2.72
2	HP	evening	45	25.0	48.8	0	0	235	2.79	8.26
3	BO	morning	20	73.5	320.5	0	0	1435	4.47	20.00
4	BO	evening	20	17.0	73.7	0	0	330	4.47	20.00

The Weibull distribution is a continuous probability distribution. The probability function $F(X)$ of a three parameter Weibull random variable X is given with the formula (9), where X_0 is the shift parameter, A is the scale parameter, and B is the shape parameter.

$$F(X) = 1 - \exp(-((X - X_0)/A)^B); \quad X \geq X_0; \quad A > 0; \quad B > 0. \qquad (9)$$

It is easy to notice that for the $B = 1$ the equation (9) represents the exponential distribution of random variable X with the mean value equal to the sum of the shift and the scale parameters, equal to $X_0 + A$. Moreover, for the $B = 1$ the log-rank transformation of the X leads to the linear regression between the $F(X)$ and the log-rank of X, given random sample of Xs. It seems, that on the explanatory analyses stage [19], it can be quite enough to put all trust on the log-rank probability plots methodology [21], avoiding the false impressions caused by inspection limited to the traditional histograms only [2, 22].

Tab. 4. The log-rank plots for concentration of the fungi in the patients bedrooms

ID	Ward	N	Log-rank equation	R^2
1	HP	90	Log-rank $= 16.58N_C + 32.824$	0.992
2	BO	40	Log-rank $= 1.510N_C + 34.508$	0.838
2	BO	40	Log-rank $= -0.67N_C^2 + 4.19N_C + 33.16$	0.978

N_c – cumulativenumber of cases, related to successive log-rank of concentration of fungi;
R^2 – coefficient of determination of the estimated log-rank equation.

The log-rank test compares area under curve (AUC) under estimates of the hazard functions for two or more groups along with all diapason from the first to the last observed event [48]. For the Weibull distribution, defined with formula (9), the hazard function is defined with formula (10).

$$h(x) = f(X)/(1 - F(X)) = (B/A) \cdot ((X - X_0)/A)^{B-1} \tag{10}$$

where: X_0 is the shift parameter, A is the scale parameter, and B is the shape parameter.

For the shape parameter equal to $B = 1$ the hazard function is stable and it is equal to reciprocal of a mean value, so the log-rang test can be considered as an exact alternative for other statistical tests applied for testing equality of the mean values [33].

Results obtained with the log-rank test

The log-rank test for differences between morning and evening concentration separately at the each wards under study, showed that both differences were insignificant here, $p = 0.26$ for the HP ward, and $p = 0.68$ for the BO ward. Therefore, it was decided to join the morning and evening data.

[Tab. 5] showed the significance equal to p(chi2) = 0.05; that is just on the borderline. This undecided result of the long-rank test corresponded to the result of comparing the frequencies of the departures over norm 50 CFU/m^3 for fungi concentration in patients' rooms, see [Tab. 6].

Tab. 5. The log-rank test for difference between mean concentration of the fungi

frequency	HP	BO	p(chi2)
observed	90	40	0.05
expected	99.6	30.4	

Tab. 6. Frequency of departures over norm for fungi concentration

ward	time	N	$N\| < 50$	$N\| > 50$	$\%\| > 50$	95%CI	
HP	morning	45	34	11	24.4%	14.2%	38.9%
HP	evening	45	38	7	15.6%	7.5%	29.2%
BO	morning	20	19	1	5.0%	0.0%	25.7%
BO	evening	20	19	1	5.0%	0.0%	25.7%

$N| > 50$; $\%| > 50$ – number (percentage) of departures over norm in total number of N events;
95%CI – confidence interval for $\%| > 50$.

Conclusions and discussion

In this study the highly skewed data from the pilot study on concentration of the fungi in the hospital wards create basis to illustrate the way for searching after appropriate nonparametric statistical procedure aimed to make comparisons between mean values. In general, the thesis that the usefulness of any nonparametric statistical procedures should be confirmed thoroughly with regard to the data distribution, was confirmed with the use of the extremely simple, but clear examples of the inappropriate understanding the aftermaths of the Box-Cox transformation to normality, and then the essence of Mann-Whitney rank-sum test. The log-rank approach was examined using the real-life data sets, obtained at two chosen hospitals. It was demonstrated, once again, that the Box-Cox transformation method can lead to erroneous conclusion even with respect to ideal log-normal populations. Then, it was demonstrated, also once again, that the results of the Mann-Whitney rank-sum test often doesn't correspond to relations neither between medians nor between means. The practical criterions, helpful to recognize the situations allowing to conclude on relations between mean values, were recommended. The real-life data sets, obtained at two chosen hospitals, corresponded to the Weibull distribution. Therefore, the log-rank approach was the primary candidate for the most acceptable method in the matter. For the shape parameter near to $B = 1$ the conclusions from the log-rank test are valid directly to relation between the mean values in the populations under investigation. Moreover, the log-rank plots and log-rank test applied together may provide a deeper insight into essentials of the investigated relationship, than the simple comparisons of the mean values only. For this reasons, the log-rang plots and the log-rank test applied jointly, seem to be quite sufficient to provide trustworthy conclusions about the aptness of the anti-fungal everyday activity in the hospital settings. Thus, the search for the most acceptable method of statistical analysis was shut in this pilot study on the log-rank approach. In case of need, other methodology should be applied with the aim to disclose the causes of the detected insufficiency of the anti-fungal practice, but it lies beyond the scope of this paper.

The present study, as each pilot study, has its typical limitations. Only two hospital wards, and the moderate number of data, $N = 2 \cdot (45 + 20) = 130$, were investigated. Nevertheless, with regard to planning further investigations, this pilot study gave rather decisive support to estimate the sufficient number of the data at each ward under study, near to these applied here, between $N = 2 \cdot 20 = 40$ and $N = 2 \cdot 45 = 90$.

Acknowledgement

The initial version of this paper was presented at the annual summer seminar of the Polish National Group of the International Society for Clinical Biostatistics (ISCB) in Krakow (Poland) on 29.06.2012. Authors are sincerely grateful to all participants in the discussion for providing their inspiring comments to our presentation, what helped us to make this paper much concise and persuasive. However, all views expressed are our.

REFERENCES

[1] Martin M. A, Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties, Computational Statistics & Data Analysis, 51, pp. 6321–6342, 2007.

[2] Callaert H., Nonparametric hypotheses for the two-sample location problem, Journal of Statistics Education, 7 (2), 1999.

[3] Lee E. T., Desu M. M., Gehan E. A., A Monte-Carlo study of the power of some two-sample tests, Biometrika, 62, pp. 425–532, 1975.

[4] Batanero C., Godino J.D., Vallecillos A., et al., Errors and difficulties in understanding elementary statistical concepts, International Journal of Mathematical Education in Science and Technology, 25 (4), pp. 527–547, 1996.

[5] Castro A. E., Vanhoof S. S., Van den Noortgate W., et al., How confident are students in their misconceptions about hypothesis tests?, Journal of Statistics Education, 17 (2), 2009.

[6] Tan W. Y., Tabatabai M. A., Some Monte Carlo studies on the comparison of several means under heteroscedasticity and robustness with respect to departure from normality, Biometrical Journal, 28 (7), pp. 801–814, 1986.

[7] Cressie N. A. C., Whitford H. J., How to use the two sample t-test, Biometrical Journal, 2, pp. 131–148, 1986.

[8] Altman D., Egger M., Gotzsche P., et al., The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies, Plods Med., 4 (10), pp. 296, 2007.

[9] Tyszka T., Pułapki psychologiczne, Psychologia biznesu dla menedżerów, Akademia Leona Koźmińskiego w Warszawie, Warszawa, 2012.

[10] Smallbone T., Quinton S., Increasing business students' confidence in questioning the validity and reliability of their research, Electronic Journal of Business Research Methods, 2 (2), pp. 153–162, 2004.

[11] Lavy L., Mashiach-Eizenberg M., The interplay between spoken language and informal definitions of statistical concepts, Journal of Statistics Education, 17 (1), 2009.

[12] Lovett M. C., Greenhouse J. B., Applying cognitive theory to statistics instruction, The American Statistician, 54 (3), pp. 196–206, 2000.

[13] Chance B. L., Components of statistical thinking and implications for instruction and assessment, Journal of Statistics Education, 10 (3), 2002.

[14] Brewer J. K., Behavioral statistics textbooks: source of myths and misconceptions?, Journal of Educational Statistics, 10 (3), pp. 252–268, 1985.

[15] Krzanowski W., Statistical principles and techniques in scientific and social research, Oxford University Press, New York, 2007.

[16] Barr J., Gould M., Joffe A., Pitfalls in the interpretation of multivariable models in the critical care literature, Chest, 127 (1), pp. 411–412, 2005.

[17] Good P. I., Hardin J. W., Common errors in statistics (and how to avoid them), Wiley Interscience, 2003.

[18] Campbell M., Machin D., Medical statistics. A commonsense approach, John Wiley & Sons, England, 1999.

[19] Behrens J. T., Principles and procedures of exploratory data analysis, Psychological Methods, 2 (2), pp. 131–160, 1997.

[20] Marchette D. J., Jeffrey L., Solka J. L., Using data images for outlier detection, Computational Statistics & Data Analysis, 43 (4), pp. 541–552, 2003.

[21] ReliaSoft Corporation, Probability plotting papers. ©1996–2006. ReliaSoft Corporation, http://www.weibull.com/GPaper/index.htm.

[22] von Hippel P. T., Mean, median, and skew: correcting a textbook rule, Journal of Statistics Education, 13 (2), 2005.

[23] Kelman B., Robbins C., Swenson L., et al., Risk from inhaled mycotoxins in indoor office and residential environments, International Journal of Toxicology, 23 (1), pp. 3–10, 2004.

[24] Maschmeyer G., Prevention of mould infections, Journal of Antimicrobial Chemotherapy, 63, (Suppl. 1), pp. i27–i30, 2009.

[25] Bodey G. P., Freireich E. J., Influence of high-efficiency particulate air filtration on mortality and fungal infection: a rebuttal, The Journal of Infectious Diseases, 194, pp. 1621–1622, 2006.

[26] Pawińska A., Mikrobiologiczne monitorowanie środowiska szpitalnego. In: Profilaktyka zakażeń szpitalnych – bezpieczeństwo środowiska szpitalnego, A. Pawińska, (Ed.), pp. 57–87, α-medicapress, Bielsko-Biała, Poland, 2011.

[27] Cimoca G., A simple algorithm for comparing hospital units efficiency, Appl Med Inform, 8 (1–2), pp. 3–7, 2001.

[28] Campbell M., Fitzpatrick R., Haines A., et al., Framework for design and evaluation of complex interventions to improve health, British Medical Journal, 321 (7262), pp. 694–696, 2000.

[29] Gniadek A., Cytotoxicity of Aspergillus fungi as a potential infectious threat. In: Insight and Control of Infectious Disease in Global Scenario, edited by. Priti Kumar Roy, InTech Rijeka, Croatia, pp. 231–248, 2012.

[30] Gniadek A., Macura A. B., Intensive care unit environment contamination with fungi, Advances in Medical Sciences, 57, pp. 283–287, 2007.

[31] Gniadek A., Skawińska M., Szczypczyk M., et al., Stosowanie klimatyzacji a występowanie grzybów w powietrzu sal bloku operacyjnego, Mikologia Lekarska, 12 (1), pp. 31–36, 2005.

[32] Knoth S., Schmid W., Monitoring the mean and the variance of a stationary process, Statistica Neerlandica, 56 (1), pp. 77–100, 2002.

[33] Watthanacheewakul L., Comparisons of power of parametric and nonparametric test for testing means of several weibull populations, Proceedings of the International MultiConference of Engineers and Computer Scientists IMECS Hong-Kong, vol. II, pp. 1534–1538, 2011.

[34] StatSoft Inc. Electronic Textbook StatSoft, Glossary, Item: Lognormal Distribution. © Copyright StatSoft, Inc., 1984–2011. http://www.statsoft.com/textbook/statistics-glossary/w/?button=0#LognormalDistribution.

[35] Chmiel I., Górkiewicz M., The bootstrap and multiple comparisons procedures as remedy on doubts about correctness of ANOVA results, Applied Medical Informatics, 30 (1), pp. 9–15, 2012.

[36] Górkiewicz M., Using propensity score with receiver operating characteristics (ROC) and bootstrap to evaluate effect size in observational studies, Biocybernetics and Biomedical Engineering, 29 (4), pp. 41–61, 2009.

[37] Van Der Laan P., Verdooren L. R., Classical analysis of variance methods and nonparametric counterpart, Biometrical Journal, 29 (6), pp. 635–655, 1987.

[38] Guo J. H., Luh W. M., Transformation works for non-normality? On one-sample transformation trimmed t methods, British Journal of Mathematical and Statistical Psychology, 54, pp. 227–236, 2001.

[39] Guo J. H., Luh W. M., An invertible transformation two-sample trimmed t statistic under heterogeneity and nonnormality, Statistics and Probability Letters, 49, pp. 1–7, 2000.
[40] Bellera C. A., Julien M., Hanley J. A., Normal approximations to the distributions of the Wilcoxon statistics: Accurate to What N? Graphical Insights, Journal of Statistics Education, 18 (2), 2010.
[41] Bereziewicz W., Górkiewicz M., Jak dużo a priori w a posteriori: poznanie naukowe z zastosowaniem metod statystyki, Cogitatum, 2, pp. 1–9, 2012.
[42] Zimmerman D. W., Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances, Journal of Experimental Education, 55, pp. 171–174, 1987.
[43] Hilgers R., On the Wilcoxon-Mann-Whitney-test as nonparametric analogue an extension of t-test, Biometrical Journal, 24 (1), pp. 3–15, 1982.
[44] Davidson R., MacKinnon J.G., Bootstrap tests: how many bootstraps?, Econometric Rev, 19, pp. 55–68, 2000.
[45] Martis M. S., Validation of simulation based models: A theoretical Outlook, The Electronic Journal of Business Research Methods, 4 (1), pp. 39–46, 2006.
[46] Aksenov S., Confidence Intervals by Bootstrap, Wolfram Research Inc., 2002. http://library.wolfram.com/infocenter/MathSource/4272/.
[47] Siniksaran R., BootStrapPackage: A Package of Bootstrap Algorithms for Mean, Simple Linear Regression Models, and Correlation Coefficient, Wolfram Research Inc., 2001. http://library.wolfram.com/infocenter/MathSource/815/.
[48] Zhou M., Log-rank Test: When does it Fail – and how to fix it, 2006. http://www.ms.uky.edu/~mai/research/LogRank2006.pdf.